Applying Time-Series Causal Discovery to Understand Algal Bloom Mechanisms

Ayush Prasad
NatureDots

Snehal Verma
NatureDots

Mohammad Aatish Khan
NatureDots

firstname.lastname@naturedots.com

Abstract

Harmful algal blooms present a significant ecological and public health challenge, yet their drivers are embedded in complex, time-lagged systems with unobserved confounding variables. Traditional correlational methods often fail to uncover the true causal structure from observational time-series data. In this work, we apply a suite of modern constraint-based causal discovery algorithms to a real-world dataset of daily water quality and weather from Lake Mendota. By explicitly modeling time lags and allowing for the presence of latent confounders using the Time-Series Iterative Causal Discovery (TS-ICD) algorithm, our approach uncovers a scientifically plausible causal graph. We demonstrate the necessity of this approach by comparing its structural output and computational complexity against a baseline algorithm (PC) that assumes causal sufficiency. Our results identify key drivers of chlorophyll-a and reveal evidence of confounding between same-day water temperature and algae levels.

1 Introduction

Harmful algal blooms (HABs) are a growing global concern, posing threats to aquatic ecosystems, local economies, and public health. Studies show that the frequency of HABs is increasing in diverse environments, from high-latitude coastal waters affected by warming and freshening (8) to inland temperate lakes like our study site, Lake Mendota, which has a long history of eutrophication driven by nutrient runoff (1). Understanding the causal mechanisms that trigger these blooms is crucial. While factors like nutrient loading, temperature, and sunlight are known drivers, their interactions are complex, involving significant time delays and confounding from unmeasured variables.

Analyzing observational time-series data from such systems is a significant challenge. While modern predictive approaches, such as physics-guided machine learning, have shown success in forecasting water quality dynamics (5), they are primarily focused on prediction rather than discovery of the underlying causal structure. Early causal approaches in Earth sciences often relied on methods like Granger causality or Dynamic Bayesian Networks (DBNs) (4), which can fail to distinguish direct from indirect causes and are highly susceptible to being misled by latent confounders (2).

To address these limitations, we apply modern causal discovery techniques to the Lake Mendota dataset (10). Our primary method is the Time-Series Iterative Causal Discovery (TS-ICD) algorithm (6), which is specifically designed for time-series data and is robust to the presence of latent confounders. By comparing its output to simpler algorithms like PC (9), we demonstrate the critical importance of selecting a method whose assumptions align with the realities of the system under study.

Our contributions are:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1. We present an application of a recently developed suite of causal discovery algorithms to a real-world limnological dataset, overcoming limitations of prior causal approaches.
2. We conduct a comparative analysis showing that explicitly modeling latent variables and leveraging temporal structure yields more efficient and plausible causal models.
3. The resulting causal graph provides new, testable hypotheses about the specific time-lagged drivers of algal blooms in Lake Mendota.

2 Methodology

We developed our methodology to discover and compare causal structures from observational time-series data. We first describe our dataset and pre-processing steps, then provide a detailed overview of the three constraint-based causal discovery algorithms used in our comparative analysis, also listing their different underlying assumptions.

2.1 Data and Pre-processing

We use a high-frequency, daily resolution dataset of water quality and meteorology for Lake Mendota, USA, sourced from the LakeBeD-US benchmark dataset (10). For our analysis, we selected a focused set of five variables known to be relevant to algal blooms: `precipitation_sum`, `temperature_2m_mean`, `shortwave_radiation_sum`, water `temp`, and `chla_rfu` (chlorophyll-a). To account for seasonality as a potential confounder, we engineered cyclical features from the date, `doy_sin` and `doy_cos`. To model time-lagged relationships, we transformed the data into a static format using a lag window of 4 days, meaning an outcome at time t is predicted by variables from $t-1$ through $t-4$. All conditional independence (CI) queries were performed using a partial correlation test (`CondIndepParCorr`) with a significance level of $\alpha = 0.01$.

2.2 Causal Discovery Algorithms

The PC (Peter-Clark) algorithm serves as our baseline, operating under the strong assumption of *causal sufficiency* (no latent confounders) to recover the causal skeleton and equivalence class (CPDAG) of the underlying graph (9). It begins with a fully connected undirected graph and first performs an adjacency search, iteratively removing edges between variables X and Y if a separating set S is found such that $(X \perp Y|S)$. After establishing the skeleton, it orients edges by first identifying v-structures ($X \rightarrow Z \leftarrow Y$) where the separating set S_{XY} does not contain the collider Z, and then exhaustively applying a set of orientation rules.

Our primary method is the Time-Series Iterative Causal Discovery (TS-ICD) algorithm, designed for time-series and *not assuming causal sufficiency*. TS-ICD outputs a Partial Ancestral Graph (PAG), which can represent confounding via bidirected edges ($\leftrightarrow$). The algorithm initializes a fully connected PAG with uncertain endpoints ($o-o$) and proceeds iteratively, performing adjacency removal and orientation phases. Critically, TS-ICD leverages the temporal structure by enforcing the arrow of time and assuming stationarity, significantly pruning the search space.

We also evaluate the Ordered Iterative Causal Discovery (OrdICD) algorithm to demonstrate the benefit of incorporating background knowledge. OrdICD also allows for latent confounders but gains efficiency by using a known causal order ($\prec$). Given the natural temporal ordering in our unrolled data, where variables at $t-k$ precede those at $t-j$ for $k > j$, the algorithm constrains its search for a separating set, which dramatically reduces the number of required CI tests.

3 Results and Analysis

Our experimental analysis is presented in two parts. First, we evaluate the computational performance of the three causal discovery algorithms. Second, we present and interpret the learned causal graph for Lake Mendota from our primary method, TS-ICD, and contrast its structure with the expected output from the baseline.

3.1 Computational Performance

We measured the computational complexity of each algorithm by the total number of CI tests required. The results (Table 1) show a clear performance hierarchy. The baseline PC algorithm was the most exhaustive (2625 tests), while OrdICD (1091 tests) and especially TS-ICD (634 tests) were significantly more efficient.

Table 1: Computational cost of causal discovery methods measured by the number of conditional independence (CI) tests. k is the conditioning order (size of the conditioning set). PC performs the most tests. ICD-Ord reduces tests by using background temporal order: TS-ICD requires the fewest especially at higher orders ($k \geq 2$).

Algorithm	**Total CI Tests**	**k=0**	**k=1**	**k=2**
PC	2625	276	1168	1181
OrdICD	1091	276	671	144
TS-ICD	634	276	319	39

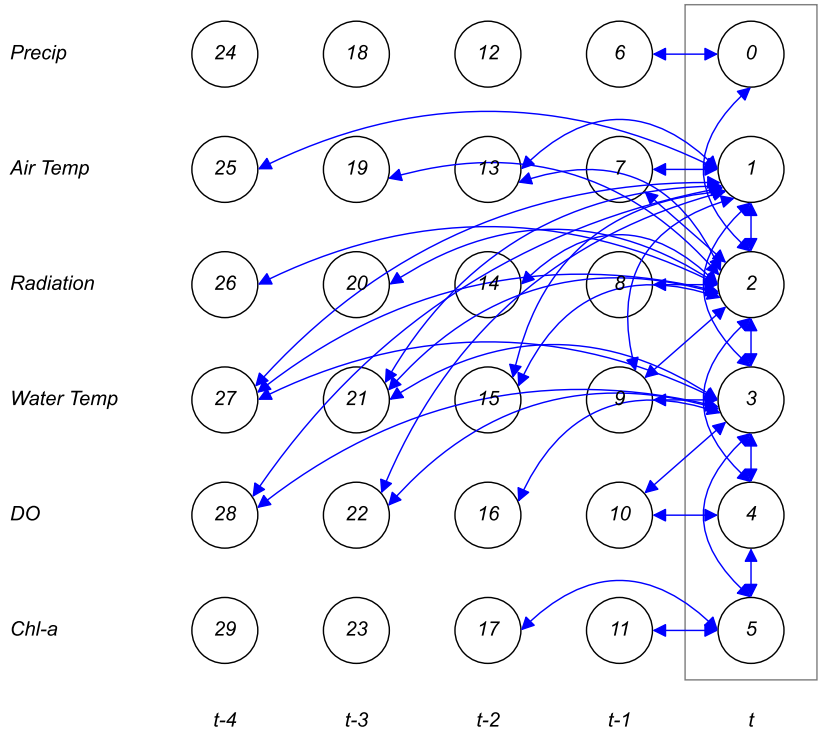

Figure 1: Partial Ancestral Graph (PAG) learned by TS-ICD for Lake Mendota. Columns show lags $t-4$ to t; rows list variables (Precip, Air Temp, Radiation, Water Temp, DO, Chl-a). Solid arrows ($\rightarrow$) denote putative direct causal effects; bidirected edges ($\leftrightarrow$) at t indicate possible latent confounding. From the graph we infer: (i) state persistence ($\text{Chl-a}_{t-1} \rightarrow \text{Chl-a}_t$ and $\text{Water Temp}_{t-1} \rightarrow \text{Water Temp}_t$); (ii) a lagged light effect $\textit{Radiation}_{t-1} \rightarrow \text{Chl-a}_t$ consistent with diel growth; (iii) a meteorological–physical link $\textit{Air Temp}_{t-1} \rightarrow \textit{Water Temp}_t$; and (iv) a contemporaneous $\text{Water Temp}_t \leftrightarrow \text{Chl-a}_t$ dependence, suggesting an unobserved regime (e.g., calm, sunny conditions) influencing both.

3.2 Causal Structure of Algal Bloom Drivers

The PAG learned by the TS-ICD algorithm for Lake Mendota (Figure 1) summarizes the inferred relationships among drivers and chlorophyll-a. This model provides a data-driven hypothesis of the dynamics driving chlorophyll-a and is consistent with basic lake physics and phytoplankton ecology. First, the graph recovers state persistence (e.g., $\mathrm{Chl\text{-}a}_{t-1} \to \mathrm{Chl\text{-}a}_t$ and $\mathrm{temp}_{t-1} \to \mathrm{temp}_t$), which is expected because biomass accumulates over multiple days and the surface layer has thermal inertia. Second, we observe a direct, time-lagged link from shortwave radiation at $t-1$ to chlorophyll-a at t. At a daily resolution this lag is plausible: light drives photosynthesis within hours, but changes in measured chlorophyll (via pigments and cell division) typically appear after one diel cycle. The graph also shows the meteorological–physical pathway in which air temperature at $t-1$ points to water temperature at t, reflecting surface heat fluxes that warm and stabilize the mixed layer.

The remaining meteorological drivers act in ways that agree with known mechanisms. Edges from wind speed point into physical variables and, where identified, into chlorophyll. Stronger winds deepen the mixed layer, dilute surface biomass, and can lower near-surface fluorescence on short timescales; they can also resuspend nutrients that influence growth with a delay. Precipitation links appear as upstream inputs consistent with runoff and nutrient pulses. Where orientation is weak at daily resolution, the PAG leaves an open endpoint, which correctly records that the data do not compel a single direction.

Most importantly, the PAG contains a bidirected edge between same-day water temperature and chlorophyll-a ($\mathrm{temp}_t \leftrightarrow \mathrm{Chl\text{-}a}_t$), indicating a latent common cause. This is a plausible representation of weather regimes that are only partially measured in our covariates. Calm, sunny conditions can simultaneously warm the surface layer and increase light exposure by reducing mixing, producing a same-day association between temperature and chlorophyll that is not purely causal in either direction. By allowing a bidirected mark, TS-ICD attributes this dependence to an unobserved factor rather than forcing a direct arrow.

This interpretation also explains the difference to the PC baseline. Because PC assumes causal sufficiency, it must explain the same-day temperature–chlorophyll correlation with a directed edge, which is likely spurious in the presence of unmeasured regime variables. TS-ICD respects temporal ordering, uses only admissible past variables in separating sets, and reports uncertainty through PAG edge marks.

4 Conclusion

In this work, we presented an exploratory study on the application of modern time-series causal discovery algorithms to a real-world observational dataset. Our comparative analysis showed that the specialized TS-ICD algorithm, which accounts for latent confounders and temporal structure, was significantly more computationally efficient than baseline methods. The resulting causal model identified plausible time-lagged drivers and, crucially, highlighted a confounded relationship between two contemporaneous variables that was missed by the baseline PC algorithm. The linear assumption of the partial correlation test is a simplification; future work could explore the use of non-parametric or kernel-based conditional independence tests to capture potentially nonlinear relationships. Our analysis also relied on an assumption of stationarity. A more advanced approach would be to investigate methods that can accommodate non-stationary or regime-shifting dynamics in the data. Furthermore, the output of our analysis is a qualitative causal graph. A valuable next step could also be to use the learned Partial Ancestral Graph (PAG) as a structural prior to fit a parametric model, such as a Structural Vector Autoregression, to quantify the strengths of the discovered causal effects. This would also provide the foundation for performing counterfactual analysis.

References

[1] S. R. Carpenter, J. J. Cole, M. L. Pace, et al. Trophic cascades, nutrient trading, and lake eutrophication. *Ecosystems*, 11(5):781–796, 2008.

[2] Michael Eichler. Causal inference in time series analysis. In *Handbook of Causal Analysis for Social Research*, pages 327–354. Springer, 2013.

[3] A. Gerhardus and J. Runge. High-recall causal discovery for autocorrelated time series with latent confounders. In *Advances in Neural Information Processing Systems*, 33, 2020.

[4] Jaroslav Hlinka, D. Hartman, M. Vejmelka, et al. Reliability of causal discovery from subsampled time series. In *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(12):126708, 2017.

[5] X. Jia, J. S. Read, J. P. Zwart, et al. Physics-guided machine learning for predicting lake water quality. In *Communications Earth & Environment*, 6(1):1–11, 2025.

[6] Raanan Yehezkel Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. From temporal to contemporaneous iterative causal discovery in the presence of latent confounders. In *International Conference on Machine Learning*, pages 39939–39950. PMLR, 2023.

[7] Raanan Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. Iterative causal discovery in the possible presence of latent confounders and selection bias. In *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Edson Silva, François Counillon, Julien Brajard, et al. Warming and freshening coastal waters impact harmful algal bloom frequency in high latitudes. In *Communications Earth & Environment*, 6(1):445, 2025.

[9] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

[10] Y. Yabansu and P. G. Hanson. LakeBeD-US: a benchmark dataset of lake water quality and meteorology for 32 US lakes. In *Earth System Science Data*, 17, 3141–3157, 2025.